7

34737

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G06F 19/00

A1

(11) International Publication Number: **WO 98/13781**

(43) International Publication Date: 2 April 1998 (02.04.98)

(21) International Application Number: PCT/US97/17201

(22) International Filing Date: 25 September 1997 (25.09.97)

(30) Priority Data:
08/741,866 26 September 1996 (26.09.96) US

(71) Applicant: PRESIDENT AND FELLOWS OF HARVARD COLLEGE [US/US]; 17 Quincy Street, Cambridge, MA 02138 (US).

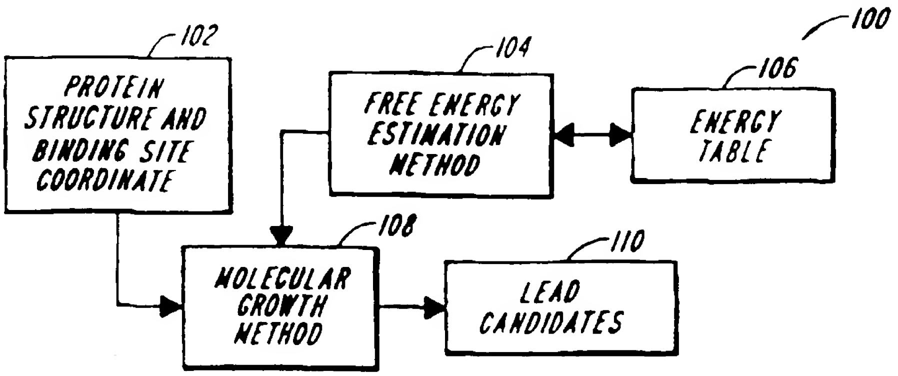
(72) Inventors: SHAKHNOVICH, Eugene, I.; 16 Woodcliffe Road, Lexington, MA 02173 (US). DeWITTE, Robert, S.; 28 Green Street, Wakefield, MA 01880 (US).

(74) Agents: KENNARD, Wayne, M. et al.; Hale and Dorr LLP, 60 State Street, Boston, MA 02109 (US).

(81) Designated States: JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: SYSTEM AND METHOD FOR STRUCTURE-BASED DRUG DESIGN THAT INCLUDES ACCURATE PREDICTION OF BINDING FREE ENERGY


100
102
PROTEIN STRUCTURE AND BINDING SITE COORDINATE
104
FREE ENERGY ESTIMATION METHOD
106
ENERGY TABLE
108
MOLECULAR GROWTH METHOD
110
LEAD CANDIDATES

(57) Abstract

A system and method for providing improved de novo structure based drug design that include a method for more accurately predicting binding free energy. The system and method use a coarse graining model with corresponding knowledge based potential data to grow candidate molecules or ligands (108). In light of the present invention using the coarse graining model, the novel growth method (108) of the present invention uses a metropolis Monte Carlo selection process (218) which result in a low energy structure that is not necessarily the lowest energy structure, yet a better candidate (110) can result.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

SYSTEM AND METHOD FOR STRUCTURE-BASED DRUG DESIGN THAT INCLUDES ACCURATE PREDICTION OF BINDING FREE ENERGY

Field Of The Invention

The present invention generally relates to systems and methods for *de novo* structure-based drug design. More specifically, the present invention relates to systems and methods for *de novo* structure-based drug design that includes a method for accurately predicting the binding free energy in building novel therapeutic molecules or ligands.

Background Of The Invention

Finding leads and optimizing the leads that are found are the goals in the development of molecules or ligands that are useful in combating disease and other abnormal body conditions. In recent years, this has involved using, among other things, a molecular similarity method of drug design. This is just one of a number of methods for determining useful leads in the search for new or improved bioactive, therapeutic molecules or ligands. The molecular similarity method of drug design is based on evaluating a large range of chemical structures to find those that show either similarity with each other or complementarity with a biochemical target structure. Chemical structures identified in this manner, usually have a reasonably high probability of binding at the target structure. This search for appropriate chemical structures is enhanced and improved by computer technology which uses search algorithms to search large databases of chemical structures.

Other methods to develop desired leads or ligands has been to use high throughput screening or combinatorial chemistry. These conventional methods have been proven, in certain circumstances, to derive desired leads and ligands.

Structure-based molecular design is yet another method to identify lead molecules for drug design. This method is based on the premise that good inhibitors possess significant structural and chemical complementarity with their target receptors. This design method can create molecules with specific properties that make them conducive for binding to the target site. The molecular structures that are designed by the structure-based design process are meant to interact with biochemical targets, for example, whose three-dimensional structures are known.

The structure-based drug design method has advantages over previous methods. One important advantage is that the structure-based drug design method provides over traditional methods of lead discovery and optimization is an awareness of the intermolecular interactions that are possible.

One goal of the structure-based drug design method is to identify lead molecules. This may be accomplished in a variety of ways. However, the principal generic steps in most structure-based drug design methods are: (1) to identify and determine the structure of the receptor site; (2) to use theoretical principles and experimental data to propose a series of putative ligands that will bind to the receptor sites (which ligands are synthesized and tested for their complementarity); (3) to make a determination of the structure of the receptor/ligand complexes that are successful in binding at low free energy levels; and (4) to iterate steps 2 and 3 in an effort to further enhance binding.

The success of a structure-based drug design method, as can be surmised, may be enhanced through the use of advanced methods of computation which will expedite the identification of key molecular fragments (which may then joined to form molecules) or whole molecules (either from a database of existing compounds or through a molecular growth algorithm). These computational advances enhance the ability to develop molecules or ligands which will successfully bind to macromolecular receptor sites.

One such advance, computational docking of molecules, takes into account the structure of a receptor site and an estimation of the binding affinity of the potential drug design candidate for the macromolecular receptor. This serves as a basis for structure-based drug design through three-dimensional databases of synthesized molecules. There are considerable mathematical challenges that must be addressed and conquered in tackling docking. These include the absence of a *clearly best* method to mathematically describe the molecular shape of the receptor site and ligand, the packing of the irregular objects together (the receptor and ligand surfaces which are to be joined), and the search issues that are related to graph theory, namely, the study of isomorphous subgraphs to make the matches of a receptor site and ligand.

Docking of ligands has three components: (1) site/ligand description, (2) juxtaposition of the ligand and the site frames of reference, and (3) the evaluation of the complementarity. With regard to the site/ligand description, the atomic coordinates of the receptor macromolecules are obtained through methods, such as X-ray crystallography, nuclear magnetic resonance (NMR), or homology modeling. Site descriptions simply may be the atomic coordinates of the receptor site, however, some notion of the chemical properties of the atoms is needed if there are hopes to measure chemical complementarity, and not just spatial complementarity. Site volume also may be defined if a site boundary is not identifiable. Ligand descriptions closely parallel site descriptions.

Structures for a large number of organic and inorganic compounds have been determined experimentally using X-ray and neutron crystallography technology. If such structures are not so defined, the approximate three-dimensional coordinates for the associated compounds may be obtained by using commercially available computer software products that use a combination of force fields and heuristic rules to generate atomic coordinates.

In considering the juxtaposition of the ligands and receptor sites, the general desire in molecular docking is to obtain the lowest free energy structure for the receptor-ligand complex. Moreover, this search for the lowest free energy selects the best fit at each position of the structure with regard to the lowest binding free energy and selection criteria. To attempt to achieve this result (1) databases of putative ligands are searched and identified ligands are ranked according to their respective interactive energies with a particular receptor site and (2) computational studies are made of the geometry of particular complexes.

In evaluating the complementarity of a receptor site and ligand, of interest for ligand design is the free energy of binding ($\Delta g_{binding}$). This may be calculated directly using free energy perturbation methods, if an accurate geometric model of the ligand-receptor complex is available. Free energy calculations generally require a great deal of computational time, *i.e.*, in the order of days. In light of this, there have been many proposed simplifications for calculating free energy; however, these simplifications have tended to add to the inaccuracy of the free energy determination. This degradation of the accuracy is unacceptable, and, as such, is there is a demand for a method to compute free energy in what is considered a reasonable amount of time, which is in the minute or less range, and accuracy does not suffer.

The calculation or prediction of free energy can be effected by any of a number of commercially available software programs. A few of these programs and their computational strategies will be discussed in greater detail subsequently.

It is known to use one of two methods to automatically search databases that contain large amounts of data relating to fragments that can be used for building molecules or ligands for developing lead candidates. A first method is the *Geometric* method that matches ligand and receptor site descriptors. A second method is to align the ligand and receptor by minimizing the ligand and receptor interactive energy. To fulfill the requirements of this method, energy-driven searching may be based on molecular dynamics ("MD") and traditional Monte Carlo ("MC") simulations. These methods, however, require a tremendous amount of computational time. Finding the lowest energy state of a given ligand-receptor complex using either of these methods is a fundamental problem.

Attempts to find ways around the actual calculation of free-binding energy has resulted in search methods based on descriptors, grids, and fragments. Looking to ***descriptor matching methods***, an analysis is made of the proposed receptor region at which binding is to take place. Ligand atoms are then positioned at the best locations at the site. This gives an approximated ligand-receptor configuration that may then be refined by optimization. Descriptor matching methods are reasonably fast and provide a good sampling of the region of interest at the receptor site. Many of the descriptor matching methods use algorithms that employ combinational search strategies. As such, small changes in parameter values can cause the computational time required to become unreasonable long.

DOCK is one of the earliest descriptor matching programs. DOCK software was developed at the University of California at San Francisco and provides a method that attempts to solve the problem by developing a drug by creating a negative image of the target site, searching a database for a similar ligand, placing putative ligands into the site, and evaluating the quality of the fit.

In using the DOCK software, an important factor is the accurate prediction of binding free energy. In favorable cases, the accuracy can reach as high as 1-2 kcal/mol. In operation, DOCK uses spheres locally complementary to the receptor surface to create a space-filling negative image of the receptor site. Several ligand atoms are matched with receptor spheres to generate chiral orientations of the ligand in the site. Databases of small molecules are searched for candidates that complement the structure of the receptor site.

There are problems with DOCK which result in predictions that are, in fact, not as accurate as desired. Further, DOCK is unable to suggest any novel structures, it can only search for what is in a database.

CAVEAT software, also a descriptor matching method, is based on directional characterization of ligands. CAVEAT was developed at the University of California at Berkley. This program searches for ligands with atoms located along specified vectors. The vectors are derived from structural information from known complexes. CAVEAT focuses on searching ligand databases to find templates as starting points for chemical structures.

FOUNDATION software provides a descriptor matching method that attempts to combine models of crucial ligand atoms and structure-based models. In using FOUNDATION, the searcher identifies atom and binding types that the candidate molecule must possess. FOUNDATION relies heavily on the detailed atom-type, bond-type, chain length, and topology constraints provided by the searcher to restrict the search. FOUNDATION only considers the steric component of the active site and relies on the matching information to find chemically complementary ligands. The tight constraints that are required by the FOUNDATION program restrict the candidates to one orientation at the receptor site.

CLIX software, developed at CSIRO, an Australian company, resembles DOCK by using the receptor site to define possible binding configurations. CLIX relies on an elaborate chemical description of the receptor site. This program uses fewer receptor-ligand matches than does DOCK. CLIX also evaluates interaction energies at the receptor sites.

Grid search methods are used to sample the six degrees of freedom of the orientation space. These methods identify an approximate solution, which cannot be guaranteed with discrete sampling methods. Accuracy is limited by the step size used in the search of the various positions. The size of step also determines the time of the search, *i.e.*, the greater the number of incremental steps, the greater the search time. Methods that use additional sampling in regions of high complementarity can overcome this problem.

A first type of grid search method is a side chain spheres method. This method explores protein-protein complexes using simplified sphere representations of side chain atoms and a grid search of four rigid degrees of freedom. This program uses surface evaluation algorithms, full molecular force-field evaluations of complexes, and simulated annealing to refine initial docking structures.

A second type of grid search method is a soft docking method. According to this method, receptor and ligand surfaces are divided into cubes to generate the translational part of the search. A pure rotational grid search is conducted on the sample ligand at orientations in discreet angular increments. The accuracy is limited by size. Run-time scaling is the cube of the rotational step size and is a product of the number of the receptor-ligand surface points.

Fragment-joining methods identify regions of high complementarity by docking functional groups independently into receptors. These methods are not particularly bothered by rigid ligand issues because of the added combinational search. Fragment-joining methods suggest unsynthesized compounds, but connecting the fragments in sensible, synthetically accessible patterns is difficult. Fragment-joining methods have the problem that there is a need to connect functional groups to form complete molecules while maintaining the fragments at the geometric positions of lowest energy.

GROW is a fragment-joining method which has been used to design peptides complementary to proteins of a known structure. The software was developed at Upjohn Laboratories, Kalamazoo, Michigan. In operation, a seed amino acid is placed in the receptor site followed by iterative additions of amino acids. Conformations are chosen from a library of precalculated low-energy forms. At each addition of a peptide, the peptide-receptor complex is minimized and evaluated. Only the best 10-100 low energy structures are kept at any stage.

HOOK, developed at Harvard University, Cambridge, Massachusetts, is a fragment-joining method that finds *hot spots* in receptor sites by looking for low energy locations for functional groups. HOOK uses random placement of many copies of several functional fragments followed by molecular dynamics.

Multiple Start Monte Carlo methods also have been used as fragment-joining methods. These methods conduct searches of databases for fragments of a ligand to dock at the receptor site.

BUILDER software, uses a family of docked structures to provide an irregular lattice of controllable density. This lattice can be searched for paths that link molecular fragments.

LUDI is a fragment-joining method that proposes inhibitors by connecting fragments that dock into microsites on the receptor. LUDI was developed at BASF, Stuttgard, Germany. The fragments are from a predetermined list of molecular fragments. The microsites are defined by hydrogen bonding and hydrophobic groups. Ligand pseudoatom positions are generated within microsites on the basis of an appropriate angle and distance minima for various interactions. The fragments identified are connected using linear chains composed of one or more of 12 functional groups.

GRID software is a hybrid grid/fragment-joining method that places small fragment probes at many regularly spaced grid points within an active receptor site. This program, developed at Oxford University, England, has been found to reproduce the positions of important hydrogen bonding groups. GRID uses empirical hydrogen bonding interaction potential and spherical representations of functional groups to generate affinity contours for various molecular fragments. This identifies regions of high and low affinity. The contours may be used to guide chemical intuition or as an input for other analysis programs. GRID is limited by its representation of the fragments since it does not allow prediction of fragment orientation.

To obtain accurate results from the programs discussed above, there is a need to understand, evaluate, and predict (or calculate) the binding free energy at the receptor sites. There also is a need to do so in a reasonable amount of computational time. The current predictive and short cut calculation methods leave much to be desired for improving structure-based drug design.

The prediction of binding free energy is according to the Expression:

$$\begin{aligned}\Delta g_{binding} &= \Delta e_{binding} - T\Delta s_{binding} \\ &= \Delta e_{complex-formation} - T\Delta s_{complex-formation} \\ &\quad + \Delta e_{solvation-desolvation} - T\Delta s_{solvation-desolvation}\end{aligned} \tag{1}$$

where,

$\Delta g_{binding}$ = The total binding free energy for a protein-ligand complex.

$\Delta e_{binding}$ = The interaction energy minus any intramolecular strain induced on complex formation.

$\Delta s_{binding}$ = The change in conformational freedom induced by the formation of the complex.

$\Delta e_{complex\ formation}$ = The interaction energy for the ligand/protein complex.

$T\Delta s_{complex\ formation}$= The change in entropy due to the reduction in flexibility in both the ligand and the protein upon complex formation.

$\Delta e_{solvation\text{-}desolvation}$ = The energies of solvation are the energetic factors from the transfer of hydrophilic and lipophilic groups from an aqueous solvent to the more lipophilic region of the protein binding site.

$\Delta s_{solvation\text{-}desolvation}$ = The entropy of solvation relating to the changes in the order of the solvent at the interface between ligand and solvent, and the protein and solvent on complex formation.

The ability to rapidly and accurately compute binding affinity of ligands for macromolecular receptors remains a problem in *de novo* structure-based drug design. Even though free energy perturbation calculations can produce relative binding free energies to within 1 kcal/mol, these methods are limited. The CPU ("Central Processing Unit") time required to ensure convergence is excessive even using the fastest computers and the method is confined to small mutations between that pair of molecules for which there is an attempt to compute the difference in binding free energy.

To emphasize this point, in *de novo* structure-based drug design, there is the need to be able to test as many molecules as possible at proposed receptor sites in a short period of time and rank the tested structures based on an accurate prediction of the binding free energy. If small organic molecules are thought of as simple combinations of functional groups, candidate molecules may be considered as a choice and an arrangement, for example, of five functional groups. As such, a database of fragments that is being used to construct such a small organic molecules can result in an extremely large number of structures that must be evaluated. For example, even a very small database of 50 fragments will produce nearly 1 billion candidates of 5 functional groups -- 50^5 combinations. As the size of the database of molecular fragments increases, it is readily seen that the number of possible combinations will increase dramatically. As such, the computational time necessary to test every combination, using current technology, is unreasonable and, therefore, not practical. The present invention provides a system and method that enhances the ability to conduct searches. In particular, a quick and accurate free energy estimation method is used for testing only the most meaningful combinations.

Conventional computational methods, such as those already discussed, use algorithms that hopefully overcome some of the computational burdens. However, in attempting to overcome the computational burdens, accuracy in predicting the binding free energy, which is a very important parameter to know in ranking possible candidates, has suffered greatly. For accurate estimates of binding free energy, sophisticated simulations using empirical potential and stepwise estimation of the changes in enthalpy and entropy at each stage of the thermodynamic cycle have been used to calculate the free binding energy. These calculations require extensive computational time for each molecule, which is impractical for normal structure-based drug design. These conventional methods, therefore, use scoring techniques that provide a short list of possible candidates for complete thermodynamic determination, or chemical synthesis and experimental determination of binding free energy.

In many of the methods for approximating the full expression for binding free energy, the scoring is based solely on the interaction energy between the ligand and protein in the complex as the single most important contributor to free energy. In others, ranking may be based more on spacial complementarity than chemical complementarity. In these cases, the solvation contributions are taken as being an approximation of the surface area. In either of these two methods the scoring is incomplete, which adversely affects the accuracy of the rankings of the candidate molecules or ligands.

In *de novo* structure-based drug design, there have been a number of methods to try to generate and rank lead candidates for docking at receptor sites. In the past, however, accuracy in predicting free energy suffered when short-cut methods were used to determine binding free energy. Therefore, unless excessive computational time was used to calculate the binding free energy in a three-dimensional protein-ligand complex, there will be a great chance of inaccuracy. Since the ranking of candidates molecules and ligands is primarily based on binding free energy, the chance of improperly ranking the candidates is very high.

The computational time needed for sampling a large number and variations of structures to generate potential lead candidates is usually very large and unreasonable. Since it is highly desirable to sample as many structures in as short a period of time as possible and then rank them based on accurate prediction of their binding free energy, there has been a need for a system and method to grow lead molecules without the computational burden of the past but with greater accuracy in binding free energy prediction.

Summary Of The Invention

The present invention is a system and method for computational *de novo* structure-based drug design that employs a novel method for discovery and building ligands, and a more accurate method for predicting binding free energy. Accordingly, the system and method of the present invention provide a better predictive *de novo* structure-based drug design tool by using a coarse-graining model with corresponding knowledge-based potential data. Moreover, in light of the use of the coarse-gaining model, the novel molecular growth method of the present invention uses a metropolis Monte Carlo selection process for molecule growth that builds the molecules or ligands that result in a low free energy structure, but not necessarily the lowest free energy structure, yet such a structure that is grown has a more accurate prediction of binding free energy and can be an acceptable drug design candidate.

The molecular growth method of the present invention uses the Metropolis Monte Carlo method to quickly search and sample the configuration space of the binding site. The is done with knowledge of the interactive potential for the fragments that are part of a database. The Metropolis Monte Carlo method also gives the system and method of the present invention the ability to identify very quickly fragments that will be useful in building the molecules or ligands.

The coarse-graining model with knowledge-based potential data follows from the application of the principles of canonical statistical mechanics to subsets of proteins, in particular the determination that small subsets of a folded protein are in thermal equilibrium with each other. As such, the present invention employs these principles. Thus, there is a reasonable basis to contend that the information present in the crystal structures of proteins and crystal structures of protein-ligand complexes may be disassembled into constituent parts and the contribution of each part to the binding free energy may be assigned on the basis of probability. The permits the present invention to achieve the more accurate results for binding free energy predictions and apply them in building the molecules or ligands.

According to the system and method of the present invention, the identification of candidate molecules is not just a search for the lowest free energy complex, but is a search to identify candidate molecules or ligands that form low free energy complexes at the receptor site but give the best lead for drug design. This is accomplished using the growth method that employs a metropolis Monte Carlo selection process. The molecular growth method results in low free energy candidates generated of a desired length.

An object of the present invention is to provide a novel system and method for structure-based drug design.

Another object of the present invention is to provide a novel system and method for structure-based drug design that has a more accurate method for predicting the binding free energy at the protein-ligand complex.

A yet further object of the present invention is to provide a novel system and method for building candidate molecules or ligands for binding at a receptor site that uses a more accurate method for predicting binding free energy at the protein-ligand complex.

These and other objects will be described in greater detail in the remainder of the specification referring to the drawings.

Brief Description Of The Drawings

Figure 1 shows a block diagram of the method employed by the system of the present invention.

Figure 2 is a block diagram for the method of estimating binding free energy that may be used in the molecule growth method employed by the system of the present invention.

Figure 3 is a flow diagram of the molecular growth method of the system of the present invention.

Figure 4 shows an example of a protein receptor site with at least an H_2 molecule loaded in it.

Figure 5 shows an example of a protein receptor site with a molecule being built in it.

Detailed Description Of The Invention

The present invention is a system and method for structure-based drug design that uses a novel method for building candidate molecules or ligands, which employs a more accurate method to predict the binding free energy of the molecules or ligands as they are grown. The system and method of the present invention also grows candidate molecules or ligands in which there is a more accurate prediction of the binding free energy in a reasonable amount of computational time.

Unless indicated otherwise, the following definitions will apply in describing the present invention:

Binding is a physical event in which a ligand is associated with a receptor site in a stable configuration.

Docking is a computational procedure whose goal is to determine the configuration that will permit binding.

De novo structure-based drug design is meant to refer to a process of dynamically forming a molecule or ligand which is conducive to binding with a particular receptor site using knowledge of the protein structure.

Ligand is a molecule that will bind with a receptor at a specific site.

Molecule (in the context of structure-based drug design) is a structure that can be formed based on the proposed receptor site.

Interaction radius is the distance within which an interaction of protein and ligand atoms must take place.

Reference state is an artificial configuration of ligand and protein atoms in which the frequency of all interaction types is exactly average. It is best described by the average probability $\bar{p} = <p_{ij}>$.

Quasichemical approximation is a mathematical method for converting probabilities to energies based on the principles of canonical statistical mechanics.

Referring to Figure 1, generally at 100, the novel method that the system of the present invention uses for building and ranking lead candidates for *de novo* structure-based drug design, as shown. In Figure 1, molecular growth method 108 receives inputs from 102 and 104. At 104, information regarding the protein structure and its binding site is provided. With regard to the binding site, at least one coordinate is provided to identify a proposed receptor site. The free energy estimate method at 104 provides the estimate of free energy of the molecule or ligand that is being built. The free energy method receives input from energy table 106 which is developed according to Figure 3, which will be described subsequently. Once a number of molecules or ligands have been built, they are ranked at 110, usually based on their binding free energy, as lead candidates for drug design.

In a little greater detail, at 102, information about the protein structure that includes the binding site is provided to molecular growth method 108. This information includes the coordinate of the binding site, protein atom coordinates, and protein atom types in a standard Brookhaven Protein Data Bank ("PDB") format or notation.

The second input to molecular growth method 108 is the free energy estimate for the molecule being built. As stated, this free energy estimate method uses the information from the calculated free energy database developed according to Figure 3. This prediction is based on proper selection of an interaction model and reference state, selection of an appropriate interaction radius, knowledge of the atom types at issue, information with regard to known structures of protein-ligand complexes (knowledge-based potential data), and quasichemical approximations.

The molecular growth method at 108 is set forth in detail in Figure 2, generally at 200, and will be described subsequently. This method employs a Metropolis Monte Carlo ("MMC") selection process to control the acceptance of intermediate and finished molecules or ligands as conditions based on their estimated binding free energy.

Referring to Figures 1 and 2, the molecular growth method at 108 will be discussed in detail. At 106 in Figure 2, the energy table (based on Figure 3) is calculated. This table is calculated once for a given set of parameters. When the parameters are changed, the table is recalculated. For example, if the interaction radius is changed from 5Å to 3Å, the table may be recalculated. It is necessary to calculate the table at or near the beginning of the method of the present invention so that it may be accessed in building molecules or ligands.

At 202 in Figure 2, the protein with a target receptor site is loaded. This is followed by step 102 at which the information about the protein that has the target receptor is input from 108 of Figure 1. This information describes the protein structure and the binding site at issue. In describing the binding site, its relative position is given in the form of a coordinate. From this point, the molecular growth method will grow a molecule or ligand that is a simple organic molecule which consists of fragments joined with single bonds.

At 204, a hydrogen molecule ("H_2") is positioned randomly in the binding site of the protein at the coordinate. This H_2 molecule is considered to be the existing molecule to satisfy step 206. According to step 206, one of the H atoms is randomly selected to be the site of the new bond.

Step 208 is performed by selecting at random a fragment from a library of fragments. For example, a library could include the fragment set forth in Table 1:

Table 1: Fragments For Molecule Growth Method

1. amide	14. cyclohexane	27. indole	39. purine
2. amine	15. cyclohexane	28. iodide	40. pyramidine
3. carbonyl	16. cyclohexane	29. methyl	41. pyradine
4. carboxylic acid	17. cyclohexene	30. n-butyl	42. pyrrole
5. chloride	18. 1,2-dithian	32. naphalene	43. sulfate

Table 1: Fragments For Molecule Growth Method

6. cyanide	19. ethane	33. nitrile	44. sulfide
7. cyclo-octane	20. ethene	34. nitro	45. t-butyl
8. cyclo-octane(2)	21. fluoride	35. phenyl	46. tetrahydrofuranyl
9. cyclo-octane(3)	22. furan	36. phosphate	47. tetrahydrothienyl
10. cyclo-pentane	24. glucose	37. propane	48. thiophene
11. cycloheptane	26. hydroxyl	38. propene	49. trifluoromethyl

At step 210, there is the random selection of at least one H atom on the randomly selected fragment. The selected H atom from the H_2 molecule and the H atom from the fragment will form the first fragment bond for the molecule being built at the protein binding site.

At step 212, the first (and new) bond is formed between the first fragment and the remaining H atom from the H_2 molecule loaded at step 204. As this bond is formed, the two H atoms that were selected are eliminated. By following the method of the present invention, it is assured that the new bond angles and bond lengths are reasonable.

The next step is at 214 where the new fragment is oriented with respect to the bond just created and binding site by torsional rotation about the new bond so that the new fragment is properly situated at the binding site at a low energy level. Preferably in orienting the new fragment by torsional rotation about the new bond, the orientation is performed in fixed increments. These fixed increments are the smallest that will still permit reasonable computational times. Orientations that are not sterically hindered are evaluated for their free energy value at step 216. The position of the fragment or rotamer that yields low energy or the lowest energy is considered as a candidate for the molecule that is being grown.

Using the molecular growth method of the present invention, it is found that molecules or ligands of any desired size may be built. This is realized by the ability to add fragments by the displacement of H atoms and binding of the atoms that were formerly connected to the displaced H atoms. The branching effect is fully realized in that any H atom of the structure that is attached to the protein is a potential growth site.

At 216, there is the prediction of free energy for the molecule being built. The method of estimating free energy is shown in detail in Figure 3 at 300. Figure 3 is a block diagram of the method for creating or calculating the energy table that is used to provide the desired free energy information for the molecule that is being grown fragment by fragment. In order to understand what is being estimated, it is necessary to understand the factors that determine the binding free energy of molecules. The remainder of the molecular growth method will be discussed before the method of estimating binding free energy is discussed in detail.

After step 216, the molecular growth method advances to step 218. Using a coarse-graining model with knowledge-based potential data, the system and method of the present invention evaluates the combinatorial search space, (the binding site of the protein), which is a rough energy landscape, to develop and identify candidate lead molecules that have a low, not the lowest, free energy complex. The system and method of the present invention overcome the multiple minimum problem, by using a MMC selection process at step 218. In order to determine whether this fragment will be accepted as a condition, once the binding free energy has been computed for each of the orientation of the rotamer, the MMC selection process at 218 makes a comparison with respect to the energy per atom before the current growth step and after the growth step at the optimal orientation. If there is a decrease in the energy per atom with the new built step, then that orientation is accepted as a condition. If an increase in energy per atom is experienced, however, it also is accepted as a condition but with a probability defined by Example (2):

$$p = \exp\left[-\frac{\Delta g}{T}\right] \tag{2}$$

where,

p = Probability of acceptance.

Δg = $\Delta G/N$ is the change in free energy per atom (with ΔG being the free energy difference upon adding the fragment at issue and N being the total number of atoms in the ligand).

T = Temperature.

The MMC selection process according to the system and method of the present invention allows the energy per atom to increase occasionally. This is needed if a small molecule is grown into a tight steric region of the binding site and the molecule had to be grown into the solvent or other unoccupied region and only marginally interact with the protein was present. Thus, allowing such an increase may provide an opportunity for a subsequent larger decrease in the free energy of binding.

Using the MMC selection process of the present invention as an optimization method has advantages. This method creates a low energy molecule in light of the random selection of rotamers that does not lead to significantly more metropolis failures. Further, the method will indirectly result in the tightest possible steric complementarity.

The step at 220 is to determine if the molecule that has been built is large enough at this point in time. If it is, then the method is directed to step 222. At step 222, it is determined if another molecule is to be built at this same binding site of the protein. If no additional molecules are built, the method will cease generating molecules for this protein binding site and the method will go to step 230 to wait for a next protein for which a candidate molecule or ligand is to be built. If the answer at step 222 is "yes" and another molecule is to be generated, then the method is directed back to step 204 where H_2 is added to the binding site of the protein where the new molecule or ligand is to be built.

If at 220 it is determined that the molecule is not large enough, then the method of the present invention returns to step 206 where another fragment is randomly selected for the addition to the existing molecule in the described manner.

In performing the method of the system of the present invention, consideration is given to the parameters surrounding performance of the method. Some of the parameters that are considered are the temperature at which the building takes place, the nearest approach allowed between atoms when assessing steric hindrance, and the angular increment in choosing fragment rotamers.

Preferably, the temperature that is selected is one that generates the largest number of low energy structures per unit of time. The nearest approach of two atoms is a percentage of the sum of their van der Waal radii since this will provide a good correlation with the nearest approaches in the database. The selected incremental amount for torsional orientation of the fragments can be further refined to smaller increments. However, such smaller increments may result in significantly more computational time to obtain results. Given the preferred parameters a molecule of at least twenty heavy atoms can be generated in about one second

The method for estimating binding free energy will now be discussed in detail referring to Figure 3. As stated, Expression (1) is used to estimating binding free energy:

$$\begin{aligned} \Delta g_{binding} &= \Delta e_{binding} - T\Delta s_{binding} \\ &= \Delta e_{complex-formation} - T\Delta s_{complex-formation} \\ &\quad + \Delta e_{solvation-desolvation} - T\Delta s_{solvation-desolvation} \end{aligned} \tag{1}$$

Again, referring to Figure 3, in estimating the binding free energy that will be part of the energy table, the items at 302 and 304 are combined with the items at 306 to generate the statistics of atomic interactions of known protein-ligand complexes, which is shown at 310. The information at 310 forms the knowledge-based potential interaction data that is used for estimating binding free energy of particular molecule being built.

At 302, a large interaction radius for the atoms that are to be bound in the protein-ligand complex is selected. A large interaction radius is selected because it will permit the solvation entropy effects to be accounted for. The most feasible length is selected for the interaction radius.

At 304, information is provided regarding the atom types based on the different classes of possible interaction. Table 2 provides examples of information that is included at 304:

Table 2: Atom-type Interaction Data

Description	Atom Type	Code
Lipophilic Carbon Atoms	Fatty Carbon	CF
Oxygen That Accepts Hydrogen Bonds	Oxygen Acceptor	OA

Item 306, therefore, includes at least information about the atom-types that are suspected will be present at the protein-ligand complex.

Now referring to the database at 306, this database includes a listing of structures of protein-ligand complexes that are known, their coordinates, and corresponding chemical elements. This database is one that can be continually updated as more information is obtained. A sample of the information that is in this database is shown in Table 3:

Table 3: Protein-Ligand Complex Database

Protein	PDB Code
Purine Nucleoside Phosphorylase	1 PNP
Amino Acid Binding Protein	1 LST

The information from 302 and 304 when combined with information at 306 result in the knowledged-based data at 310. Specifically, the information at 310 is the statistics of atomic interactions in known protein-ligand complexes which will permit a more accurate prediction of binding free energy for the molecule or ligand being built.

The next issue is to compile the statistics at 310 into a set of interaction parameters that constitute the free energy contribution of interactions between specific atom types. Given the probability of atomic interactions in known protein-ligand complexes from 310, this information, along with reference state for the molecule or ligand being built, are combined. This reference state is selected such that it accounts for the solvent energy and configuration entropy effects.

The third item that is combined to generate the estimate from binding energy is the quasichemical approximation of the protein-ligand complex of the molecule or ligand being built that is at 312.

When the information from 308, 310, and 312 is combined, the results is the energy table at 106 that is used to provide the estimated free energy contributions for each type of interaction possible for the molecule or ligand being built. An example of the data provided in the free energy table at 106 is shown in Table 4:

Table 4: Energy Tables

Protein Atom	Ligand Atom	Δg_{ij}
CF	CF	-2.18
CF	OA	-1.13

Table 4: Energy Tables

Protein Atom	Ligand Atom	Δg_{ij}
OA	CF	-0.89

Before discussing the statistical support for the present invention, Figures 4 and 5 will be described. Those Figures graphically demonstrate aspects of Figures 1, 2, and 3. Referring to Figures 2 and 4, when protein 402, shown generally at 400, is loaded, it is defined and the coordinate of the binding site, such as coordinate "X" at 404, is determined. Next H_2 molecules, such as H_2 molecule 406, is added to the binding site. As would be understood, more than one H_2 molecule may be added to a single binding site.

After the H_2 molecule is loaded in the binding site, either H atom 408 or 410 is selected for forming the new bond. In this case, H atom 410 is selected. Once the selection is made, a fragment is randomly chosen from the fragment library and an H atom of that fragment is selected for establishing the new bond.

Referring to Figures 2 and Figure 5, generally at 500, one H atom of fragment 502 is selected for forming the first bond. When this bond is formed, the selected H atoms of H_2 molecule 406 (Figure 4) and fragment 502 are eliminated. When new bond 504 is formed, the first fragment has been added in building the molecule.

The added fragment is now incremently rotated about bond 504 to obtain the best fit and the free energy estimation is made based on the different orientations. An evaluation of the new molecule is made based on the MMC selection method described previously, then the remainder of the method of Figure 2 is carried out.

As stated, the present invention, implements a coarse-graining model with corresponding knowledge-based potential data. This model is used because of the ability to apply the principles of canonical mechanics to subsets of folded proteins that are in thermal equilibrium with one another. As such, the crystal structures of proteins and crystal structures of protein-ligand complexes may be disassembled into constituent parts. The contribution of each part may then be assigned on the basis of probability.

Generally, in statistical mechanics any two states at the same energy have an equal probability of the occupation of a location. This is expressed in Expression (3):

$$p^{e}_{ij} = \frac{\exp\left[-\frac{e_{ij}}{kT_{e}}\right]}{Z} \qquad (3)$$

where,

p^{e}_{ij} = The energetic probability of an interaction between a protein atom of type i and ligand atom of type j.

e_{ij} = The energy of the interaction of a protein atom of type *i* and ligand atom of type *j*.

k = Boltzmann Constant.

T_{e} = Experimental temperature at which the data base information is obtained.

Z = Normalization Constant.

In forming protein-ligand complexes, configurations that have exactly the same energy are not equally likely to be formed. This is because of the strong presence of a boundary in the space sampled by the ligand. The entropic effects are solvent ordering and restrictions on the atomic interactions due to steric hindrance and the fixed chemical structures of the protein and ligand. The entropic effects are not correlated to the energetic events, and, therefore, may be expressed as the sampling probability p^{s}, which relates to the entropic effects, as set forth in Expression (4):

$$p^{s}_{ij} = \frac{\exp\left[-\frac{s_{ij}}{k}\right]}{Z} \qquad (4)$$

where,

p^{s}_{ij} = The sampling probability of an interaction between a protein atom of type *i* and ligand atom of type *j*.

s_{ij} = Entropy of interactions between a protein atom of type *i* and ligand atom of type *j*.

k = Boltzmann Constant.

Z = Normalization Constant.

The product of p^e from Expression (3) and p^s from Expression (4) gives a relationship between the total probability and gross binding free energy (that is dependent on the interaction model chosen to describe the atomic interactions between the protein and ligand in the protein-ligand complex). This is shown in Expression (5):

$$p_{ij} = p^e_{ij} p^s_{ij} = \frac{\exp\left[-\frac{e_{ij} - Ts_{ij}}{kT_e}\right]}{Z} = \frac{\exp\left[-\frac{g^*_{ij}}{kT_e}\right]}{Z} \qquad (5)$$

where,

p_{ij} = The total probability of interactions between a protein of type *i* and ligand atom of type *j*.

p^e_{ij} = The energetic probability of an interaction between a protein atom of type *i* and ligand atom of type *j*.

p^s_{ij} = The sampling probability of an interaction between a protein atom of type *i* and ligand atom of type *j*.

e_{ij} = The energy of the interaction of a protein atom *i* and ligand *j*.

s_{ij} = Entropy of interactions between a protein atom of type *i* and ligand atom of type *j*.

k = Boltzmann Constant.

T_e = Experimental temperature at which the data base information is obtained.

Z = Normalization Constant.

The interaction model that is chosen is comprised of an interaction radius and a set of eligible atom types.

Expression (5) can be changed so that it is solved for g^*_{ij}, the free energy as set forth in Expression (6). This comprises the quasichemical approximation. Expression (6) is determined from the frequency of observed interactions:

$$g^*_{ij} = -kT_e \log(p_{ij}) - kT_e \log(Z) \qquad (6)$$

where,

g^*_{ij} = The gross free energy of an interaction between a protein atom of type i and ligand atom of type j.

T_e = Experimental temperature at which the data base information is obtained.

k = Boltzmann Constant.

Z = Normalization Constant.

p_{ij} = The total probability of an interaction between a protein of atom type i and ligand atom of type j.

When the appropriate reference state is selected, the Normalization Constant can be eliminated. The reference state contributes a free energy of $\bar{g}$ to every gross free energy term $g_{ij}*$. This is shown via Expressions (7), (8), and (9):

$$\bar{g} = -kT_e\log(\bar{p}) - kT_e\log(Z) \qquad (7)$$

$$g_{ij} = g^*_{ij} - \bar{g} \qquad (8)$$

$$g_{ij} = -kT_e\log\left[\frac{p_{ij}}{\bar{p}}\right] \qquad (9)$$

where for Expressions (7), (8), and (9),

$\bar{g}$ = Gross free energy of the reference state.

$\bar{p}$ = The average probability of an interaction between protein and ligand atoms.

g_{ij} = Net free energy of the interaction between a protein atom of type i and ligand of type j.

g^*_{ij} = The gross free energy of an interaction between a protein atom of type i and ligand atom of type j.

p_{ij} = The total probability of an interaction between a protein of atom type i and ligand atom of type j.

T_e = Experimental temperature at which the data base information is obtained.

k = Boltzmann Constant.

Z = Normalization Constant.

Expression (9) relates the statistical information about interatomic interactions in the crystal structures of the protein-ligand complex to term by term contributions to binding free energy. The g_{ij}'s, when summed, will approximate the complete form of free energy in Expression (1) that is provided at 104.

The correct interaction model and reference state for application of a knowledge-based potential data to a protein-ligand binding event is deducted from the terms in Expression (6) (repeated below:):

$$g^*_{ij} = -kT_c\log(p_{ij}) - kT_c\log(Z) \qquad (6)$$

More specifically, changes in the solvation entropy with regard to the complex formation occur because of a gain or loss in the solvent order, which is a correlation between the potential surface of the solvent exposed to atoms in the ligand, protein, or complex. These correlations are generally twice the size of a water molecule beyond the boundary of the ligand, protein, or complex. As the interactions between the ligand and protein result in desolvation, there is a change in the configurational entropy. Thus, to form a particular intermolecular contact between a protein and ligand, each of the atoms in contact must be desolvated. Where there has been a large amount of order destroyed by this process, there is an entropic increase due to the desolvation for the formation of that particular contact.

The selection of the large interaction radius, between the protein and ligand, should be the correlation length of solvent ordering. When this is the case, the probability of the specific contacts observed occurring will include the average effect of the contribution of solvation entropy to the prediction of free energy. As such, the system and method of the present invention will select a large interaction radius for the interaction model. For the purposes of the present invention, the interaction model will define a ligand atom to be in contact with a protein atom if there are within the interaction radius of one another.

Each contact formed involves an energy loss based on desolvation. This is must be accounted for in the reference state. In defining the reference state, the specificity of the loss due to desolvation of the specific contact becomes a general element that is factored in and the remaining energetic contributions in the interaction model with a large interaction radius take into account simply that a loss due to desolvation has taken place.

In selecting the reference state, there is effectively unrestricted spatial sampling of the ligand with respect to the protein. As such, the reference state has no perceived notion of the chemical structure. Thus, the difference between the free estimates of a specific structure and that of the reference state accounts for the loss of configurational entropy in a collection of atoms upon formation of a specific, largely rigid chemical structure. If Expression (8) is now considered

$$g_{ij} = g^*_{ij} - \bar{g} \tag{8}$$

in which $\bar{g}$ is subtracted from g^*_{ij}, the entropic effect of unrestricted sampling and the effect of desolvation is taken into account.

Using the Expressions set forth above, the system and method of the present invention score and then rank the candidate structures based, in large part, on the determination of the total binding free energy that is defined by Expression (10):

$$\Delta G = \sum_{ij} g_{ij}\Delta_{ij} \tag{10}$$

where,

ΔG = The total binding free energy.

g_{ij} = Net free energy of the interaction between a protein atom of type *i* and ligand of type *j*.

Δ_{ij} = This term is zero unless i and j are within the interaction radius of each other, in which case, it is equal to one.

Using the interaction model and reference state that was previously discussed, ΔG is an approximation to the complete change in free energy in the complex formation. Coarse-graining that was discussed included the entropic effects of solvation and the reference state has taken into consideration the effects of solvation energy and the configurational entropy. Moreover, according to the system and method of the present invention, a closer look is taken of the atom types and their affect on the energy contributions that really take place in complex formation, not a generalization of the atom types.

To test the accuracy of the approximation of binding free energy, the method of the present invention was compared with experimental measurements of binding free energy. The method for approximating binding free energy according to the present invention was applied to guanine based ligands that had been designed, synthesized, and assayed for purine nucleoside phosphorylase ("PNP"). In Table 5 below, each of the molecules were interactively built using the system and method of the present invention and low energy conformation was found with the conformational search according to the present invention. As such, the molecules were tested as if they had been generated using the *de novo* growth method of the present invention. In other words, given enough computation time the system and method of the present invention would have generated the listed molecules and any corresponding conformations, even though, undirected generation of these exact ligands is highly improbable. The set of molecules in Table 5 was generated by the system and method of the present invention for correlation between free energies taken as the logarithm of the binding constants or IC_{50} measurement, which is an experimental determination of the propensity for complex formation, and knowledge-based potentials of the present invention:

Table 5: Binding Constants vs. Invention Knowledge-Based Potentials for PNP Ligands

Ligand	Invention Knowledge-Based Potential	K_i or $IC^a{}_{50}$(μM)
GDP	-13.9	1500
GMP	-14.1	2000
GTP	-14.7	2400
dGDP	-14.9	590
dGMP	-14.4	1400
3-(trifluoromethyl)phenymethyl-9-deazaguanine	-12.3	0.036^a

Table 5: Binding Constants vs. Invention Knowledge-Based Potentials for PNP Ligands

Ligand	Invention Knowledge-Based Potential	K_i or $IC^a{}_{50}$(μM)
acyclovir	-15.8	91
pyridin-3-ylmethyl-9-deazaguanine	-18.5	0.025[a]
2-hydroxyphenylmethyl-9-deazaguanine	-18.1	0.270[a]
4-hydroxyphenylmethyl-9-deazaguanine	-18.7	0.260[a]
phenylmethyl-9-deazaguanine	-18.7	0.051[a]
2-(2-phosphatidylethyl) phenylmethylguanine	-17.8	0.035[a]
3-methylphenylmethyl-9-deazaguanine	-19.4	0.057[a]
3-methoxyphenylmethyl-9-deazaguanine	-18.1	0.082[a]

The knowledge-based potential data, according to the present invention, correlates well with the experimental binding free energy for over five orders of magnitude in the binding constants. The strong correlations that were found for the binding free energy predictions according to the system and method of the present invention indicate an ability to effect *de novo* drug design of lead molecules.

The system of the present invention may be operated in one of three modes: automatic, directed, or assisted. In the automatic mode, all that is necessary is to provide the starting protein structure and a coordinate on the protein that is the specific binding site. From that point on, the system will generated ligands with at least one atom within an interaction radius length of the specified coordinate. The directed mode is an interaction method in which the user specifies the molecular fragments which are selected and where they are to bind. The assisted mode starts with the user specifying a fragment. Then, the assisted mode proceeds automatically.

The terms and expressions which are used herein are used as terms of expression and not of limitation. There is no intention in the use of such terms and expressions of excluding the equivalents of the features shown and described, or portions thereof, it being recognized that various modifications are possible in the scope of the present invention.

Claims:

1. A system for building molecules for binding at a receptor site, comprising:

means for evaluating a receptor site for a molecular make up of at least a portion of the receptor site to which a molecule being grown will bind and generating at least a coordinate of at least the portion of the receptor site to which the molecule being grown will bind, and providing as an output, at least with respect to the molecular make up of the receptor site, the coordinate of the portion of the receptor site to which the molecule being grown will bind;

estimating means for estimating free energy of the molecule being grown, with the estimating means using knowledge-based potential data to estimate free energy and providing as an output the estimated free energy; and

molecular growth means that receives as inputs the output from the means for evaluating the receptor site and providing at least the coordinate of least a portion of the receptor site, and the output from the estimating means to grow the molecule to bind to the receptor site, with the molecular growth means building the molecule by selecting molecular fragments at orientations for building the molecule that result in free energy estimates for the molecule that may be higher than a lowest free energy estimate possible for the molecule.

2. A method for building molecules for binding at a receptor site, comprising the steps of:

(a) evaluating a receptor site for a molecular make up of at least a portion of the receptor site to which a molecule being grown will bind and generating at least a coordinate of least a portion of the receptor site to which the molecule being grown will bind, and outputting, at least with respect to the molecular make up of the receptor site, the coordinate of the portion of the receptor site to which the molecule being grown will bind;

(b) estimating free energy of the molecule being grown using knowledge-based potential data to estimate free energy and outputting the estimated free energy; and

(c) building a molecule for binding to the receptor site using the outputs from steps (a) and (b), with the building step including building the molecule by selecting molecular fragments at orientations that will result in free energy estimates for the molecule that may be higher than a lowest free energy estimate possible for the molecule

3. The method as recited in claim 2, wherein step (b) for estimating free energy includes the substep of:

(1) selecting an interaction radius of a length that negates adverse affects from solvent entropy effects;

(2) developing an accessible listing regarding known structures of molecule/receptor complexes that includes at least molecular fragment

(3) developing an accessible listing regarding atom types based on classes of molecular interactions that include at least atom types predicted to interact in the molecule/receptor complex of the molecule being grown and the receptor site;

(4) generating an accessible listing of molecular interactions for known structures of molecule/receptor complexes based on the selection at substep (1) and the listing developed at substeps (2) and (3);

(5) selecting a reference state for the molecule being grown that negates the adverse effects from solvent and configurational entropy effects;

(6) developing a quasichemical approximation for the molecule that is being grown; and

(7) generating the estimated free energy of the molecule being grown based on the listing developed at substep (4), the selection at substep (5), and the quasichemical approximation developed at substep (6).

4. The method as recited in claim 2, wherein a molecule can be grown according to steps (a) to (c) in less than 5 seconds.

5. The method for estimating free energy in a molecule being grown for connection at a receptor site in a molecule/receptor complex, comprising the steps of

(a) selecting an interaction radius of a length that negates adverse affects from solvent entropy effects;

(b) developing an accessible listing of known structures of molecule/receptor complexes that includes at least atom coordinates and corresponding chemical elements;

(c) developing an accessible listing of atom types based on classes of molecular interactions that include at least atom types predicted to interact in the molecule/receptor complex of the molecule being grown and the receptor site;

(d) generating an accessible lisung of molecular interactions for known structures of molecule/receptor complexes based on the selection at step (a) and the listing developed at steps (b) and (c);

(e) selecting a reference state for the molecule being grown that negates the adverse effects from solvent and configurational entropy effects;

(f) developing a quasichemical approximation for the molecule that is being grown; and

(g) generating the estimated free energy of the molecule being grown based on the listing developed at step (d), the selection at step (e), and the quasichemical approximation developed at step (f).

6. A method for building a molecules for binding at a receptor site, comprising the steps of:

(a) evaluating a receptor site for a molecular make up of at least a portion of the receptor site to which a molecule being grown will bind and generating at least a coordinate of least a portion of the receptor site to which the molecule to being grown will bind,

(b) loading at least hydrogen ("H") atoms at least at the portion of the receptor site to which the molecule being grown will bind;

(c) randomly selecting an H atom of the H atoms loaded at the receptor site:

(d) randomly selecting a molecular fragment from the predetermined group of molecular fragments and randomly selecting a H atom from the molecular fragment;

(e) forming a bond at the selected H atoms selected at steps (c) and (d);

(f) orienting the molecular fragment with respect to the bond formed at step (e) so that the molecule being grown has low free energy;

(g) estimating free energy for the molecule being grown using knowledge-based potential; and

(h) comparing the estimated free energy at step (g) with a free energy estimate of a molecule bound at the receptor site before the fragment was added at step (d) to determine if the estimated free energy at step (h) is more or less than a prior estimated free energy for the molecule before the fragment are added at step (d), and accepting the addition and orientation of the molecular fragment if the estimated free energy is less and accepting the addition and orientation of the first molecular fragment according to a probability determined in a predetermined manner if the estimated free energy is more.

7. The method as recited in claim 6, wherein step (g) for estimating free energy includes the substep of:

(1) selecting an interaction radius of a length that negates adverse affects from solvent entropy effects;

(2) developing an accessible listing regarding known structures of molecule/receptor complexes that includes at least atom coordinates and corresponding chemical elements;

(3) developing an accessible listing regarding atom types based on classes of molecular interactions that include at least atom types predicted to interact in the molecule/receptor complex of the molecule being grown and the receptor site;

(4) generating an accessible listing of molecular interactions for known structures of molecule/receptor complexes based on the selection at substep (1) and the listing developed at substeps (2) and (3);

(5) selecting a reference state for the molecule being grown that negates the adverse effects from solvent and configurational entropy effects;

(6) developing a quasichemical approximation for the molecule that is being grown; and

(7) generating the estimated free energy of the molecule being grown based on the listing developed at substep (4), the selection at substep (5), and the quasichemical approximation developed at substep (6).

8. The method as recited in claim 6, wherein a molecule can be grown according to steps (a) to (h) in less than 5 seconds.

9. The method as recited in claim 6, wherein orienting the molecular fragment in step (f) includes torsional rotation in fixed increments.

10. The method as recited in claim 9, where it fixed increments for torsional rotation includes fixed increments of 60°.

11. The method as recited in claim 6, wherein the method of comparing and accepting the molecule is according to a metropolis Monte Carlo selection process.

12. The method as recited in claim 6, wherein the probability in step (h) is

determined according to:

$$p = \exp\left[-\frac{\Delta g}{T}\right]$$

where,

p = Probability of acceptance.

Δg = $\Delta G/N$ is the change in free energy per atom (with ΔG being the free energy difference upon adding the fragment at issue and N being the total number of atoms in the ligand).

T = Temperature.

13. A method for building a molecules for binding at a receptor site, comprising the steps of:

(a) evaluating a receptor site for a molecular make up of at least a portion of the receptor site to which a molecule being grown will bind and generating at least a coordinate of least a portion of the receptor site to which the molecule to being grown will bind;

(b) loading at least hydrogen ("H") atoms at least at the portion of the receptor site to which the molecule being grown will bind;

(c) randomly selecting an H atom of the H atoms loading at the receptor site:

(d) randomly selecting a molecular fragment from the predetermined group of molecular fragments and randomly selecting a H atom from the molecular fragment;

(e) forming a bond at the selected H atoms selected at steps (c) and (d);

(f) orienting the molecular fragment with respect to the bond formed at step (e) so that the molecule being grown has low free energy;

(g) estimating free energy for the molecule being grown using knowledge-based potential; and

(h) comparing the estimated free energy at step (g) with a free energy estimate of a molecule bound at the receptor site before the fragment was added at step (d) to determine if the estimated free energy at step (h) is more or less than a prior estimated free energy for the molecule before the fragment are added at step (d), and accepting the addition and orientation of the molecular fragment if the estimated free energy is less and accepting the addition and orientation of the first molecular fragment according to a probability determined in a predetermined manner if the estimated free energy is more.

14. The method as recited in claim 13, wherein the method further includes repeating steps (c) to (g) for each molecular fragment that is added.

15. The method as recited in claim 13, wherein step (g) for estimating free energy includes the substep of:

(1) selecting an interaction radius of a length that negates adverse affects from solvent entropy effects;

(2) developing an accessible listing regarding known structures of molecule/receptor complexes that includes at least atom coordinates and corresponding chemical elements;

(3) developing an accessible listing regarding atom types based on classes of molecular interactions that include at least atom types predicted to interact in the molecule/receptor complex of the molecule being grown and the receptor site;

(4) generating an accessible listing of molecular interactions for known structures of molecule/receptor complexes based on the selection at substep (1) and the listing developed at substeps (2) and (3);

(5) selecting a reference state for the molecule being grown that negates the adverse effects from solvent and configurational entropy effects;

(6) developing a quasichemical approximation for the molecule that is being grown; and

(7) generating the estimated free energy of the molecule being grown based on the listing developed at substep (4), the selection at substep (5), and the quasichemical approximation developed at substep (6).

16. The method as recited in claim 13, wherein a molecule can be grown according to steps (a) to (h) in less than 5 seconds.

17. The method as recited in claim 13, wherein orienting the molecular fragment in step (g) includes torsional rotation in fixed increments.

18. The method as recited in claim 17, where it fixed increments for torsional rotation includes fixed increments of 60°.

19. The method as recited in claim 13, wherein the method of comparing and accepting the molecule is according to a metropolis Monte Carlo method.

20. The method as recited in claim 19, wherein the probability in step (g) is

determined according to:

$$p = \exp\left[-\frac{\Delta g}{T}\right]$$

where,

p	=	Probability of acceptance.
Δg	=	$\Delta G/N$ is the change in free energy per atom (with ΔG being the free energy difference upon adding the fragment at issue and N being the total number of atoms in the ligand).
T	=	Temperature.

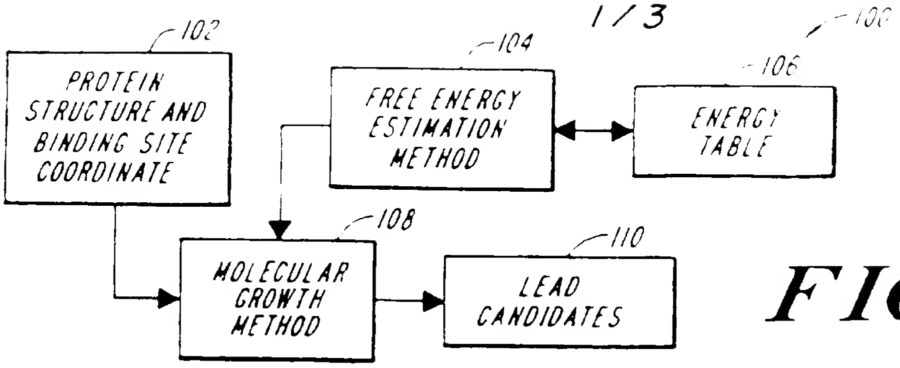

100
102
PROTEIN STRUCTURE AND BINDING SITE COORDINATE
104
FREE ENERGY ESTIMATION METHOD
106
ENERGY TABLE
108
MOLECULAR GROWTH METHOD
110
LEAD CANDIDATES

FIG. 1

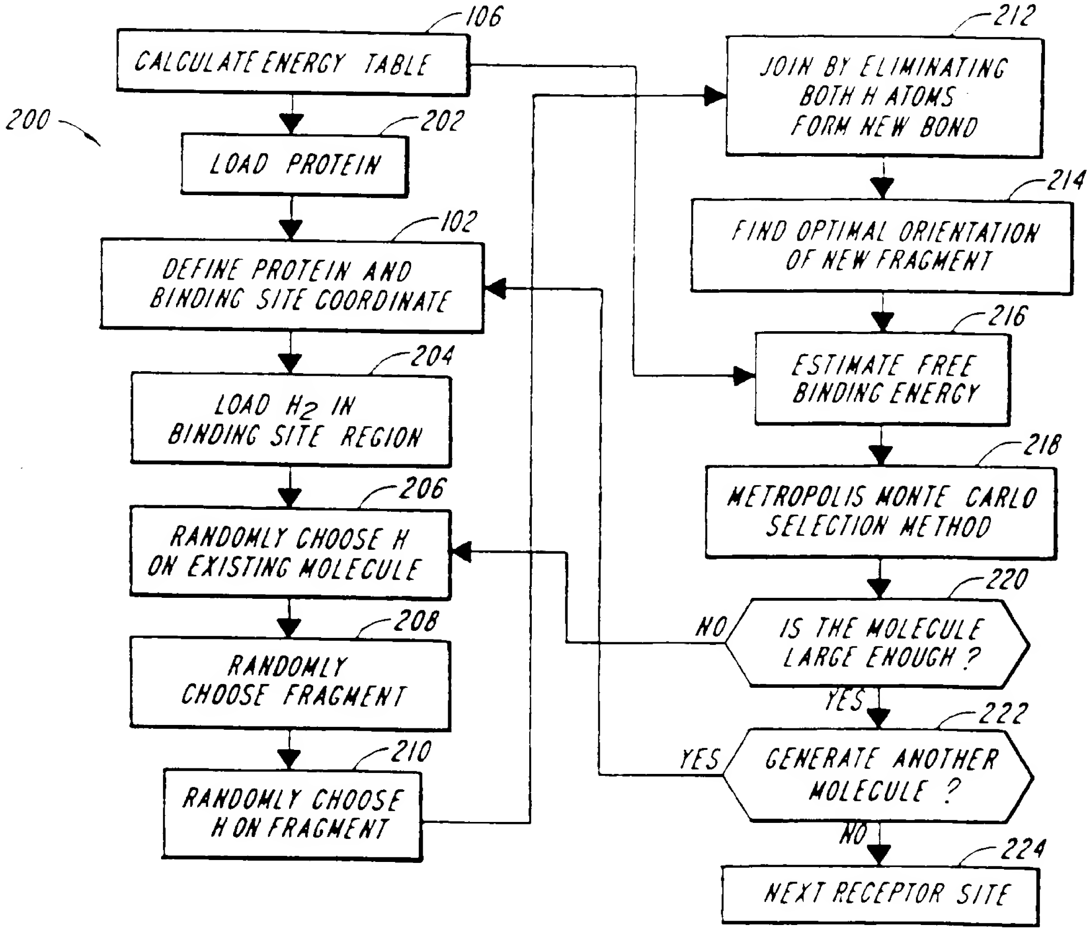

106
CALCULATE ENERGY TABLE
200
202
LOAD PROTEIN
102
DEFINE PROTEIN AND BINDING SITE COORDINATE
204
LOAD H2 IN BINDING SITE REGION
206
RANDOMLY CHOOSE H ON EXISTING MOLECULE
208
RANDOMLY CHOOSE FRAGMENT
210
RANDOMLY CHOOSE H ON FRAGMENT
212
JOIN BY ELIMINATING BOTH H ATOMS FORM NEW BOND
214
FIND OPTIMAL ORIENTATION OF NEW FRAGMENT
216
ESTIMATE FREE BINDING ENERGY
218
METROPOLIS MONTE CARLO SELECTION METHOD
220
NO
IS THE MOLECULE LARGE ENOUGH ?
YES
222
YES
GENERATE ANOTHER MOLECULE ?
NO
224
NEXT RECEPTOR SITE

FIG. 2

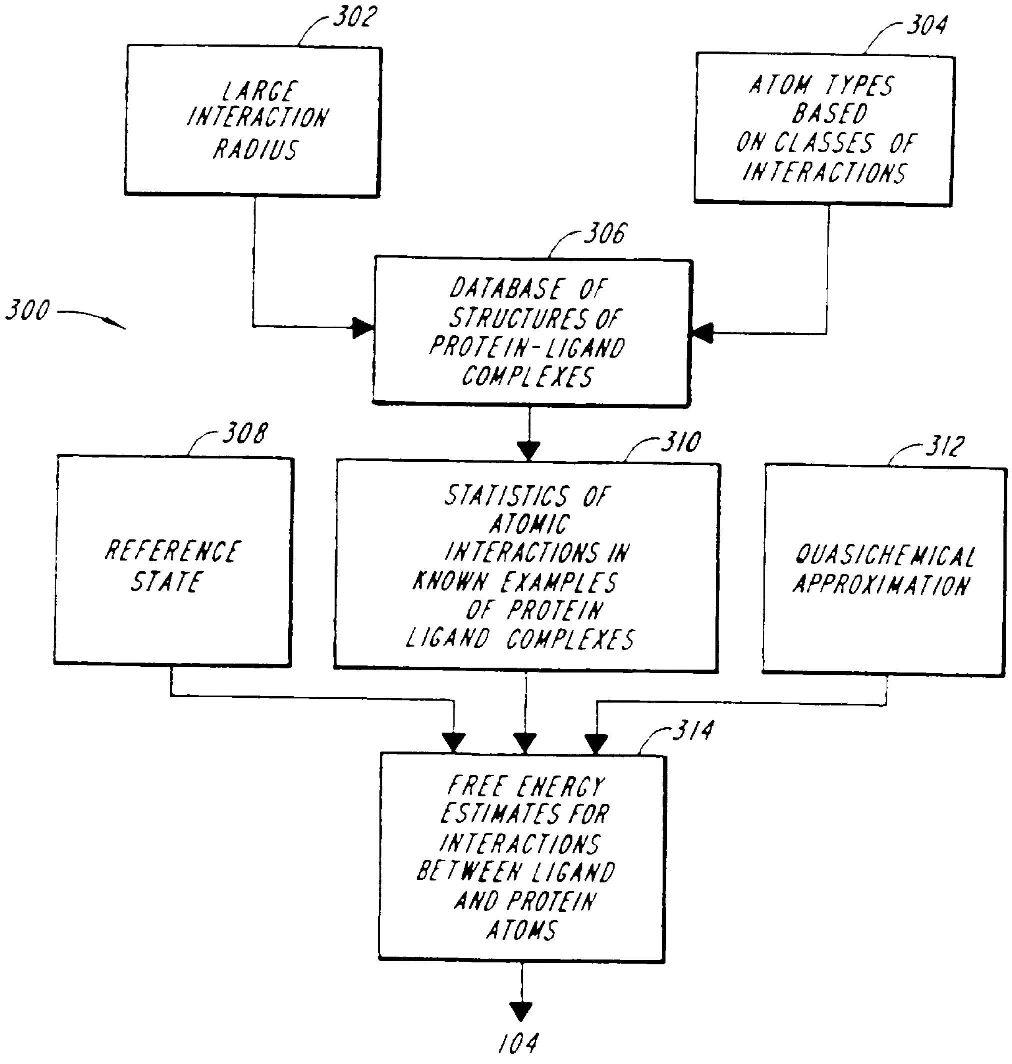

302
LARGE INTERACTION RADIUS
304
ATOM TYPES BASED ON CLASSES OF INTERACTIONS
300
306
DATABASE OF STRUCTURES OF PROTEIN-LIGAND COMPLEXES
308
REFERENCE STATE
310
STATISTICS OF ATOMIC INTERACTIONS IN KNOWN EXAMPLES OF PROTEIN LIGAND COMPLEXES
312
QUASICHEMICAL APPROXIMATION
314
FREE ENERGY ESTIMATES FOR INTERACTIONS BETWEEN LIGAND AND PROTEIN ATOMS
104

FIG. 3

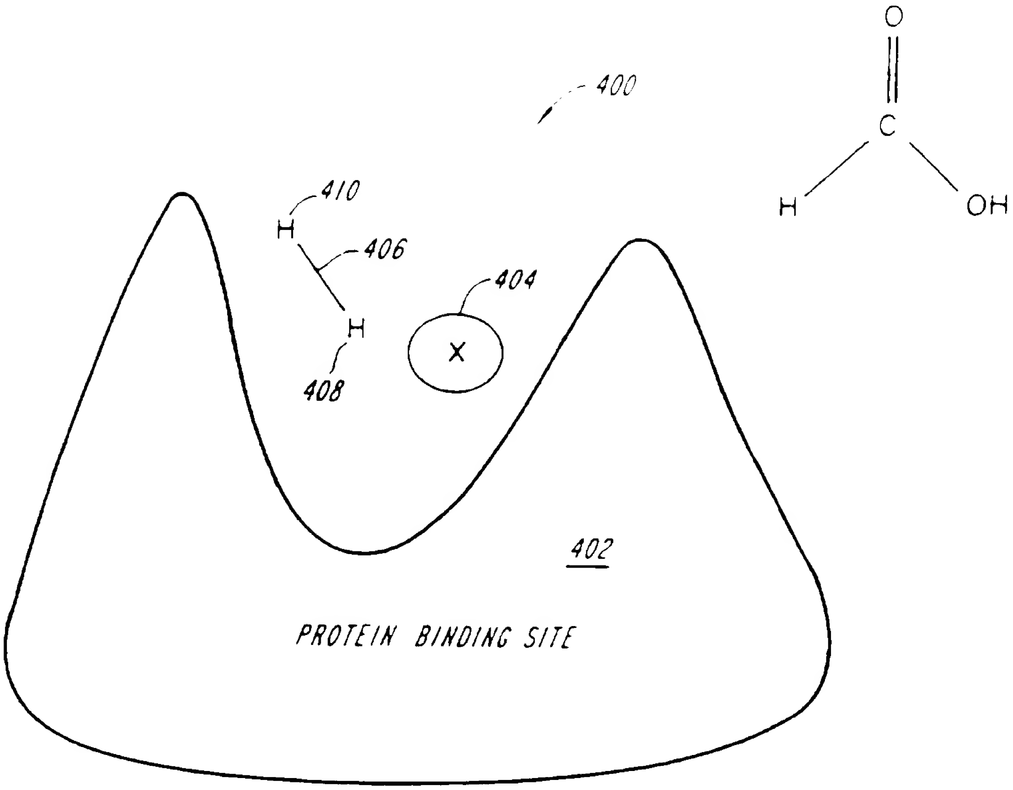

400
O
C
H
OH
410
H
406
404
H
X
408
402
PROTEIN BINDING SITE

FIG. 4

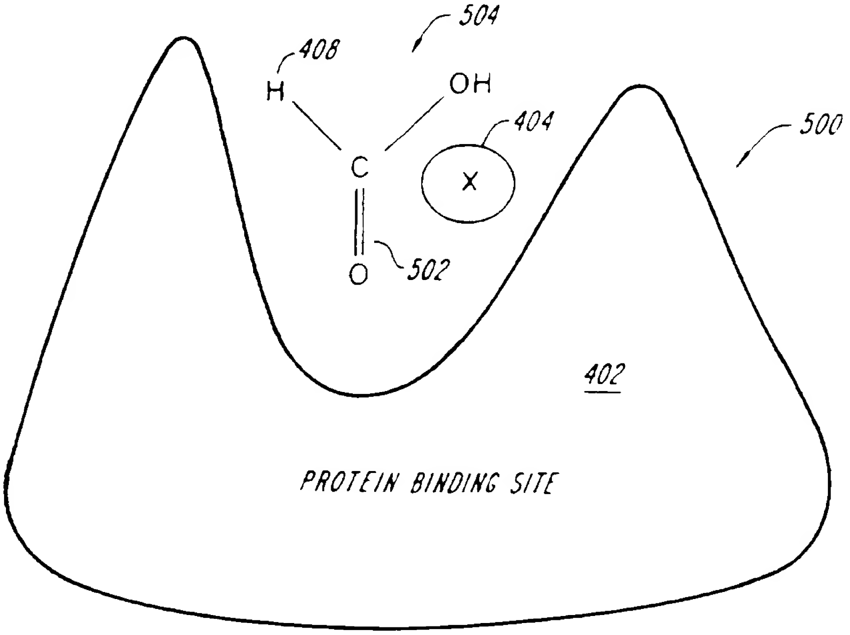

504
408
H
OH
404
500
C
X
O
502
402
PROTEIN BINDING SITE

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/US97/17201

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) G06F 19/00
US CL 364/496
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. 364/496-499, 578

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No
A	KLEBE, GERHARD. "Toward a More Efficient Handling of Conformational Flexibility in Computer-Assisted Modelling of Drug Molecules" Perspectives in Drug Discovery and Design, 1995. vol. 3, p. 85-105 especially pages 91-96	1-20
A	BOHM, GERALD. "New Approaches in Molecular Structure Prediction" Biophysical Chemistry, March 1996 vol. 59, n.1-2 p.1-32 especially 19-22	1-20
A	BLANEY, FRANK. "Molecular Modelling in the Pharmaceutical Industry" Chemistry and Industry, December 1990 n.23, p791-794 especially p.791-794	1-20

□ Further documents are listed in the continuation of Box C. □ See patent family annex.

*	Special categories of cited documents	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance, the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance, the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search

21 NOVEMBER 1997

Date of mailing of the international search report

2 8 JAN 1998

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
MELANIE KEMPER
Telephone No. (703) 305-9589